Goodness of Fit of a Markov Chain Model for Sequences of Wet and Dry Days[1]

A. M. FEYERHERM AND L. DEAN BARK

Kansas State University, Manhattan

(Manuscript received 5 April 1967, in revised form 23 May 1967)

ABSTRACT

Evidence from locations in the north central region of the United States indicates that a rainy spell is more likely to terminate after at least two wet days than after one wet day during early spring. However, the departure from a first-order Markov chain model has only minor effects when estimating probabilities of specified sequences of wet and dry days.

1. Introduction

In a previous paper, the authors (1965) developed procedures for estimating the probability of occurrence for a given sequence of consecutive wet and dry days based on the assumption that such a sequence could be adequately described as a first-order (simple) Markov chain. The works of Hopkins and Robillard (1964), Wiser (1965) and Green (1964) suggest that added precision might be obtained at some locations by assuming either higher-order Markov chain models or modifications of a simple Markov chain, or alternative models.

Previous work of the authors, which used data from six locations in the north central region, also gave evidence of imperfection when using a first-order Markov chain. Subsequently, a detailed and more exhaustive investigation was made of short-term sequences of wet and dry days to study the adequacy of a first-order Markov chain model. The results form the content of this paper.

2. Comparison of conditional probabilities

Data for this study were taken from records at the same six locations used in the previous investigation of the authors. Average annual precipitation ranged from 17.5 inches at Garden City, Kan., to 41.8 inches at Markland Dam, Ind.

Entries in Table 1 are arithmetic means of relative frequencies obtained by computing the proportion of times that a wet day (precipitation ≥ 0.01 inch) was preceded by the specified sequences of events. For the first week ($t=1, 2, \cdots, 7$) of the climatological year, the relative frequency of a wet day preceded by the sequence DW was given by the ratio

$$\frac{\sum_{y=1}^{Y}\sum_{t=1}^{7} N(D_{t-2,y}W_{t-1,y}W_{t,y})}{\sum_{y=1}^{Y}\sum_{t=1}^{7} N(D_{t-2,y}W_{t-1,y})},$$

where $N(\ )=1$ if the sequence, shown in parentheses, occurred, $N(\ )=0$ if the sequence, did not occur, and $Y=$ number of years of record. The sum of such relative frequencies, over 52 one-week periods, divided by 52, yielded a table entry.

Means shown in Table 1 mask the effect of seasonal variation. The effect is shown in Fig. 1 for selected conditional probabilities. Estimated probabilities were obtained by fitting the first four terms of a Fourier series to the observed weekly relative frequencies. The estimation procedure follows methodology outlined in the author's previous paper.

Though the entries in Table 1 mask seasonal effects, differences and similarities between specified pairs of mean relative frequencies are readily apparent. The difference of about 0.20, at all stations, between relative frequency of occurrence of a wet day following a wet day compared with a wet following a dry day reiterates the need to use at least a first-order Markov chain to estimate probabilities of sequences using a minimum number of parameters.

All average relative frequencies for a wet day immediately preceded by a dry day are nearly equal in value and there is no apparent need to consider precipitation history prior to the preceding day. However, when a wet day is preceded by a wet day, the relative frequencies vary more. Differences between the relative frequencies for a wet day preceded by the sequence WWW compared with DWW might well be ascribed to sampling variation. The differences are not in the same

[1] Contribution No. 111, Department of Statistics and Statistical Laboratory, and No. 109, Department of Physics, Kansas Agricultural Experiment Station, Kansas State University, Manhattan.

TABLE 1. Relative frequency of a wet day following a given sequence of dry and wet days. (Entries are averaged over 52 weekly values.)

Weather on preceding day(s)	Markland Dam, Ind.	Columbia, Mo.	Ames, Ia.	Columbus, Kans.	Manhatten, Kans.	Garden City, Kans.
Unknown	0.28	0.31	0.26	0.25	0.23	0.16
W	0.40	0.44	0.38	0.41	0.38	0.32
WW	0.38	0.38	0.33	0.38	0.34	0.31
DW	0.41	0.49	0.41	0.43	0.40	0.33
WWW	0.35	0.39	0.32	0.36	0.34	0.31
DWW	0.42	0.36	0.36	0.41	0.34	0.31
WDW	0.41	0.49	0.41	0.43	0.40	0.33
DDW	0.39	0.51	0.42	0.43	0.41	0.32
D	0.24	0.24	0.21	0.20	0.18	0.13
WD	0.23	0.24	0.22	0.22	0.20	0.15
DD	0.24	0.24	0.21	0.20	0.18	0.13
DWD	0.22	0.24	0.22	0.22	0.18	0.15
WWD	0.23	0.24	0.22	0.22	0.22	0.15
WDD	0.24	0.25	0.22	0.23	0.19	0.14
DDD	0.25	0.24	0.21	0.19	0.18	0.13

Note: $DW = D_{t-2}, W_{t-1}$; $DWW = D_{t-3}, W_{t-2}, W_{t-1}$; etc.

direction for all locations and when curves, similar to those in Fig. 1, were drawn to show seasonal variation, they crossed each other several times. The same was true of differences between the relative frequencies for a wet day preceded by the sequence *WDW* compared with *DDW*.

Differences between relative frequencies for a wet day preceded by the sequence *WW* compared with *DW* cannot be ascribed solely to sampling variation. For each of the six stations the yearly average relative frequency of a wet day after only one wet day is greater than the average relative frequency of a wet day following a sequence of at least two wet days. The phenomenon is examined more fully in Fig. 1, which shows estimated probabilities based on fitting the first four terms of a Fourier series to the 52 relative frequencies computed on a weekly basis.

Let

$P(W_t|W_{t-1},D_{t-2})$ = probability that the tth day is wet given that the $(t-1)$st day is wet and the $(t-2)$nd day is dry,

and let $P(W_t|W_{t-1},W_{t-2})$ be defined in a similar manner. Then, for a particular day t, the difference between estimated probabilities must exceed approximately 0.09 to reject the hypothesis

$$P(W_t|W_{t-1},W_{t-2}) = P(W_t|W_{t-1},D_{t-2})$$

at the 0.05 level. Periods when that condition is satisfied are quite numerous for Columbia but are confined to April for Garden City. The only time of the year when the results tend to agree for all locations is during the latter part of March and all of April. Fortunately, the differences tend to be smallest during the growing season when one might be particularly interested in estimating probabilities for sequencies of wet and dry days. The results agree well with similar data shown by Hopkins and Robillard (1964) for April through September at three Canadian stations in the prairie provinces.

3. Probabilities of sequences of wet and dry days

If it is true that at certain periods of the year

$$P(W_t|W_{t-1},W_{t-2}) < P(W_t|W_{t-1},D_{t-2}),$$

then it is of interest to estimate the difference between using a second-order and a first-order Markov chain model to estimate probabilities of relatively short sequences of wet and dry days. To estimate that difference, consider the week 6–12 December at Columbia, Mo., where estimated conditional probabilities were

$$\hat{P}(W_t|W_{t-1},W_{t-2}) = 0.291, \quad \hat{P}(W_t|W_{t-1},D_{t-2}) = 0.479,$$

one of the larger differences that occurred.

Estimated probabilities for all possible combinations of sequences of wet and dry days over a four-day period are shown in Table 2 assuming, in turn, independence P_0, a first-order Markov chain P_1, and a second-order Markov chain P_2 as descriptive of the succession of events. The last two columns compare (P_1-P_0) and (P_2-P_1), respectively. The differences (P_1-P_0) show sizeable errors for certain sequences if one assumed that the sequence of events are independent. However, the differences shown under (P_2-P_1) are sufficiently small that they can be ignored in many practical investigations.

Use of a second-order Markov chain model in which one has to estimate four parameters vs. two for a first-order chain, appears unnecessary, but as a compromise one could assume that

$$P(W_t|D_{t-1},D_{t-2}) = P(W_t|D_{t-1},W_{t-2}) = P(W_t|D_{t-1}).$$

In that case, only three parameters would have to be estimated with the remainder estimated by mathematical relationships that exist between the param-

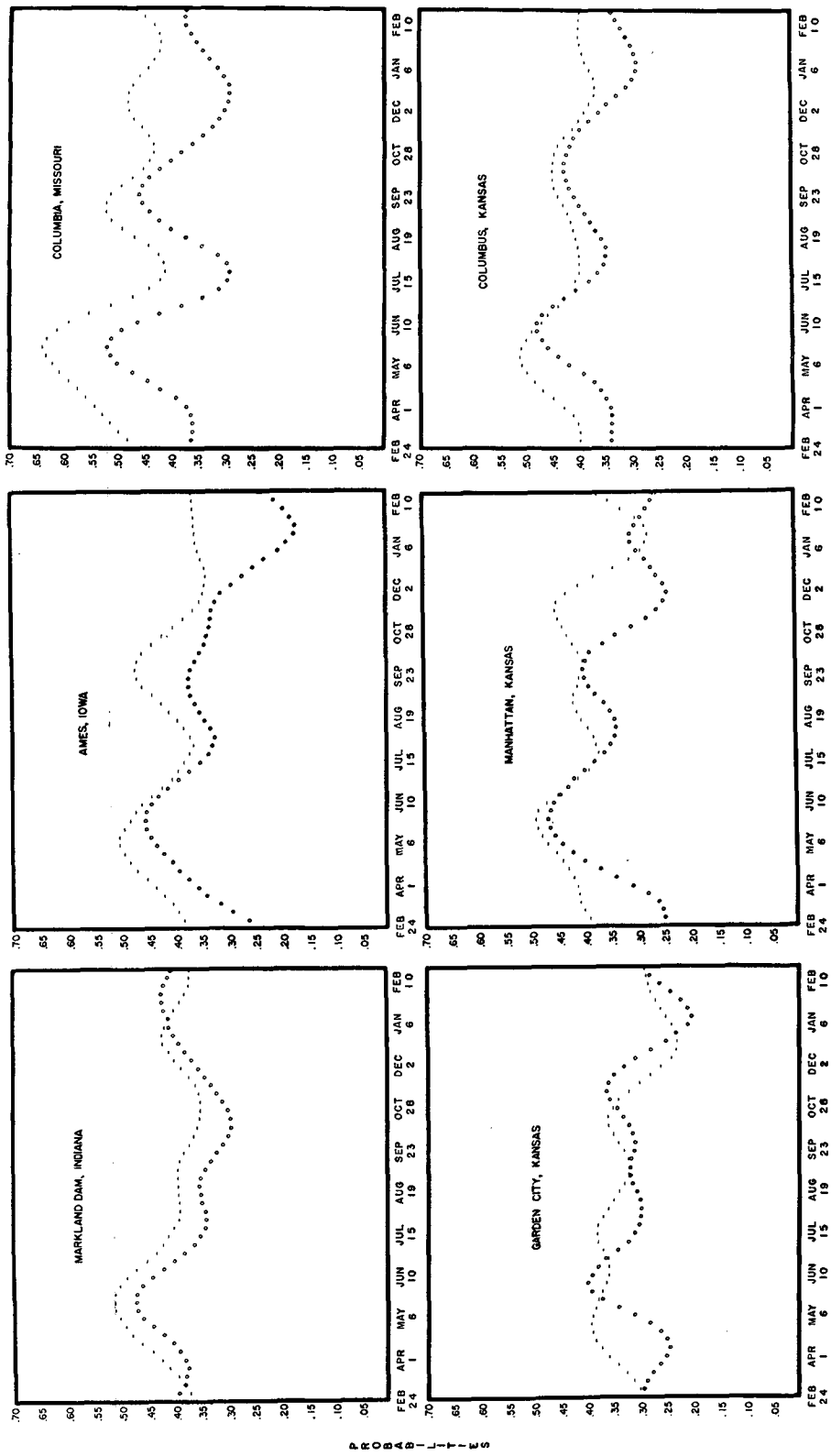

FIG. 1. Estimated conditional probabilities for a wet day $\hat{P}(W_t|W_{t-1},D_{t-2})$, dashed line; $\hat{P}(W_t|W_{t-1},W_{t-2})$, line with open circles.

TABLE 2. Estimated probabilities for four-day sequences of wet and dry days assuming various degrees of independence, Columbia, Mo., 6–12 December.

Sequences†	P_0*	P_1**	P_2***	P_1-P_0	P_2-P_1
DDDD	0.303	0.371	0.371	0.068	0.000
DDDW	0.105	0.096	0.096	−0.009	0.000
DDWD	0.105	0.072	0.063	−0.033	−0.009
DWDD	0.015	0.072	0.063	−0.033	−0.009
WDDD	0.015	0.097	0.096	−0.008	−0.001
DDWW	0.037	0.049	0.058	0.012	0.009
DWDW	0.037	0.019	0.017	−0.018	−0.002
WDDW	0.037	0.025	0.025	−0.012	0.000
DWWD	0.037	0.037	0.052	0.000	0.015
WDWD	0.037	0.019	0.017	−0.018	−0.002
WWDD	0.037	0.049	0.059	0.012	0.010
DWWW	0.013	0.025	0.021	0.012	−0.004
WDWW	0.013	0.013	0.015	0.000	0.002
WWDW	0.013	0.013	0.016	0.000	0.003
WWWD	0.013	0.025	0.022	0.012	−0.003
WWWW	0.004	0.017	0.009	0.013	−0.008

† Note: $DDDW = D_{t-3}, D_{t-2}, D_{t-1}, W_t$; etc.
* Assuming independence.
** Assuming conditional probabilities depend only on previous day's weather.
*** Assuming conditional probabilities depend on weather of two previous days.

eters. However, one must also consider that the standard error of estimate for $P(W_t|W_{t-1}W_{t-2})$ may be 1.5 to 2.0 times larger than that for $P(W_t|W_{t-1})$ for data similar to that used in this study. Hence, precision is gained by using a first-order Markov chain model at the expense of a possible increase in bias.

4. Summary and conclusions

As a follow-up to a previous investigation on the use of a first-order (simple) Markov chain model for estimating probabilities for sequences of wet and dry days, the authors considered differences that might be encountered in using a second-order in preference to a first-order chain. Statistical tests indicate that when the probability of a wet day was conditioned on weather for two previous days (second-order chain), the hypothesis

$$P(W_t|W_{t-1}W_{t-2}) = P(W_t|W_{t-1}D_{t-2}), t = 1, 2, \cdots, 365,$$

should be rejected in favor of

$$P(W_t|W_{t-1}W_{t-2}) < P(W_t|W_{t-1}D_{t-2})$$

at certain locations and for particular periods of the year—mainly the latter part of March and during April. The results agree with data from locations in the Canadian Prairie Provinces reported by Hopkins and Robillard (1964).

On the other hand, there was no substantial statistical evidence for rejecting the hypothesis

$$P(W_t|D_{t-1}D_{t-2}) = P(W_t|D_{t-1}W_{t-2}),\ t = 1, 2, \cdots, 365,$$

nor for use of a third-order in preference to a second-order Markov chain model.

The adequacy of the first-order Markov chain model for computing probabilities may not be satisfactory for long sequences, especially for prolonged dry spells when a different set of meteorological forces may be operative. In practice, where interest centers on computing probabilities for all possible sequences, the length of sequences will be relatively short. For such sequences a first-order chain appears quite adequate as indicated by an example of a four-day sequence at Columbia, Mo. Estimated probabilities for $P(W_t|W_{t-1}W_{t-2})$ and $P(W_t|W_{t-1}D_{t-2})$ differed by 0.188 but the largest difference between probabilities, when comparing a second-order and first-order chain model, was equal to 0.015 for the sequence *DWWD* and less than 0.010 for the other sequences.

Acknowledgments. The authors gratefully acknowledge computer programming services of Sai-Sing Lin and Hou-Pen Chen, graduate research assistants in the Department of Statistics, Kansas State University.

REFERENCES

Feyerherm, A. M., and L. D. Bark, 1965: Statistical methods for persistent precipitation patterns. *J. Appl. Meteor.*, 4, 320–328.

Green, J. R., 1964: A model for rainfall occurrence. *J. Roy. Stat. Soc.*, **B26**, 345–353.

Hopkins, J. W., and P. Robillard, 1964: Some statistics of daily rainfall occurrence for the Canadian prairie provinces. *J. Appl. Meteor.*, 3, 600–602.

Wiser, E. H., 1965: Modified Markov probability models of sequences of precipitation events. *Mon. Wea. Rev.*, 93, 511–516.